DOCUMENT RESUME

ED 306 268 TM 013 085

AUTHOR Johnson, Eugene
TITLE Issues and Procedures in Analyzing the State Assessment Data.
PUB DATE Mar 89
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Demography; Elementary Education; Elementary School Students; Error of Measurement; Estimation (Mathematics); Mathematical Models; National Programs; *Scaling; *Standardized Tests; *State Programs; *Statistical Analysis; *Testing Programs
IDENTIFIERS *National Assessment of Educational Progress; *Trial State Assessment

ABSTRACT

The 1990 National Assessment of Educational Progress (NAEP) will mark the first time that state-level comparisons will be possible with NAEP data. The 1990 data will contain, as a trial component, state-level assessments of eighth graders in mathematics. Mathematics and reading comparisons for fourth graders will be added in 1992. Statistical issues associated with these innovations include those related to the following: (1) estimation of within state population characteristics, including estimating variability and accounting for error of measurement; (2) comparison of the data from trial state assessment with data from the national assessment, including evaluating the effect of type of administration and equating the observed administration results to the national results; and (3) between state comparisons. Scaling models should be used to produce estimates of the distribution of abilities within the subject areas, the models should distinguish between content areas, and the models require special technology to take the amount of available information into account. Sample structure must also be addressed. It was noted that the fact that the states differ in terms of demographic, socioeconomic, and other indicators is important, and that these differences may affect the comparability of achievement results between states. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306268

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

EUGENE JOHNSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ISSUES AND PROCEDURES IN ANALYZING THE STATE ASSESSMENT DATA

Eugene Johnson
Educational Testing Service

TM013085

ISSUES AND PROCEDURES IN ANALYZING THE STATE ASSESSMENT DATA*

Eugene Johnson
Educational Testing Service

INTRODUCTION

The 1990 National Assessment of Educational Progress will mark the first time that state level comparisons will be possible with NAEP data. The 1990 National Assessment will contain as a trial component state level assessments of eighth-grade students in mathematics; the trial component will be expanded for the 1992 assessment when, in addition to state-level assessments of eighth-grade students in mathematics, there will be state-level assessments of fourth-grade students in mathematics and reading. Three purposes of the Trial State Assessments are: (1) to provide information about the performance of the students within each participating state; (2) to compare each participating state with the nation as a whole in terms of achievement results; and (3) to compare participating states to each other.

The analysis of the data from the Trial State Assessments must be such that each of these purposes can be achieved. First, to provide achievement data for each of the participating states, it is necessary to take into account the characteristics of the within state samples. The complex sample design that will be used to collect assessment data within each state will produce response data which will have different properties from those needed to apply standard statistical methodology. In order to summarize the distribution of achievement across students, scaling models should be used. Because of the design of the trial state assessment, these scaling models will produce scale scores with different properties than ordinary test scores. The first section of this paper will discuss these scaling models and explain how the structure of the data can be taken into account for analysis.

In order to compare the results of a participating state to the results of the nation, it is important to account for the potential effects due to differences in the data collection procedures to be used in the state and

---

*Paper presented as part of the NAEP symposium, "State Comparable Assessments: New Technical Challenges for NAEP, The Nation's Report Card," American Educational Research Association, 1989 Annual Meeting, San Francisco.

TM013085

national assessments. Furthermore, it is possible that motivation in the state assessments may be higher than in the national assessment due to accountability concerns. The techniques which allow the equating of the results from the state assessments to the results of the national assessment will be discussed in the second section with emphasis on the consequences of the fact that not all states will participate in the Trial State Assessments.

Finally, the third section considers the comparison of one state to another. At issue is how best to account for differences between the states demographic, socioeconomic, instructional, and attitudinal indicators in the process of comparing performance across states.

1. ESTIMATION OF WITHIN STATE POPULATION CHARACTERISTICS

A goal of the Trial State Assessment is to obtain various estimates of achievement for subpopulations of students within each of the participating states. Examples of estimates of interest are: the overall mathematics achievement for the population of students in a state who are in the eighth grade; the average score on a particular exercise; or the proportion of students in the state who have been exposed to a specific type of instruction. These estimates must come from the responses of the students who were selected for the assessment. To ensure achieved samples which allow estimates of high precision while, at the same time, being economically and operationally feasible to obtain, the Trial State Assessments will use a complex sample design.

The properties of a sample from a complex design are very different from those of a simple random sample in which every student in the target population has an equal chance of selection and in which the observations from different sampled students can be considered statistically independent of one another. These special properties must be taken into account in estimation of population characteristics. Because of the similarities between the Trial State Assessment and the National Assessment, we can infer the properties of a sample from a state assessment based on our experience with the National Assessment samples.

Like the National Assessment, the state assessment samples should have associated with them sampling weights to account for the sample design including the probabilities of selection for each student in the sample. Additionally, the weights should have adjustments for nonresponse at the student and school level. To reduce the variability of the weights and to improve the representativeness of the sample, the weights should be further adjusted, by post-stratification, to make sample estimates of certain population totals conform to external, more accurate, estimates based on information from the Census and other sources. These sampling weights should be used as required to obtain the proper estimates of all population and subpopulation characteristics.

The population and subpopulation characteristics which should be estimated within each state include the following basic information: the proportions of students responding correctly to each exercise; the proportions of students who possess various study habits, attitudes, experiences or background characteristics; and the proportions of students exposed to various educational practices. Each of these estimates should be provided for the state as a whole as well as for selected demographic subgroups of the state (such as subgroups defined by sex, race/ethnicity, or type of community under the constraint that the number of sampled students within a subgroup is sufficient to provide reliable estimates.

In addition to providing measures of the average performance of students within a state, it is important to provide information about the distribution of abilities across the students within the state. This can be achieved by applying the scaling technology developed for the National Assessment to the data from the Trial State Assessment. As has been demonstrated by NAEP, properly computed scale scores can provide a meaningful summarization of the data. For example, by using the scale scores, it is possible to relate proficiency across the subject area (or across a defined facet of the subject area) to the attitudes, background, and educational experiences of the students. Furthermore, besides providing estimates of the average proficiency of students, the NAEP scale score technology allows a description of the distribution of proficiencies across the population of students. This description, which can also be related to the background, attitudes, and

educational experiences of the students, is in terms of the proportion of students attaining each of a series of defined levels of proficiency. To aid interpretation, each of the defined levels is accompanied by a description of the types of tasks that students attaining that level of proficiency would be quite likely to successfully accomplish.

To achieve the maximum utility, the scaling should be within distinct subareas of the subject area. By scaling within narrowly defined content areas, the assumptions inherent within the scaling models are more nearly met. Furthermore, such narrowly defined subscales provide the capability of detecting interactions between the various subareas and subgroups of the population. For certain purposes, the level of detail provided by separate subscale values is too great. To fulfill the need to have a single, overall, summary measure of proficiency within a subject area, a composite score can be defined as a weighted average of the subscale scores.

In constructing the subscale and composite scores, it is important to account for the potential effects on achievement of differences in administration between the Trial State Assessment and the National Assessment. It is also important to account for the effects of measurement error in the construction of the individual level subscale scores. We will deal with the latter consideration first, in the next subsection, deferring a discussion of the former consideration until Section 2.

1.1 Estimation of Variability

Besides deriving appropriate point estimates of population characteristics, it is essential to obtain appropriate measures of the degree of uncertainty about the actual values of those statistics. There are two components of uncertainty which must be accounted for in conducting statistical analyses of data from the Trial State Assessment:

1) uncertainty due to sampling variability measuring the dependence of the results on the particular sample achieved as opposed to the results from a different sample which might have been obtained by use of the same design and field procedures, and

2) uncertainty due to errors of measurement arising because subscale estimates for each respondent are based on a relatively small number of cognitive items.

We will attend to the first uncertainty below, deferring a discussion of the second component until later in this subsection.

Because of the effects of cluster selection (first schools are selected and then students are selected within those schools) and because of effects of nonresponse and post-stratification adjustments, observations made on different students cannot be assumed to be independent of each other (and, in fact, will be positively correlated). Treatment of the data as a simple random sample, with disregard for the special characteristics of the sample design, will produce underestimates of the true sampling variability. A complex design such as that which will be used for the Trial State Assessment makes the proper estimation of the sampling variability of a statistic complicated and requires techniques beyond those commonly available in standard statistical packages.

One such technique is the jackknife procedure which has been used for the analysis of NAEP data since the inception of NAEP. The jackknife variance estimator correctly estimates the variability of a linear estimator (such as a total); produces appropriate variance estimates for nonlinear statistics such as ratios, regression estimates or weighted means; and, through the creation of replicate weights, provides for the measurement of the variability attributable to the use of poststratification and other weight adjustment factors which are dependent upon the observed sample data.

Briefly, the jackknife estimate of the variance of a statistic t is obtained by creating M pairs of first-stage sampling units and then creating a pseudoreplicate for each pair. A pseudoreplicate is an estimate of the statistic of interest t based on an altered sample. Specifically, the i^{th} pseudo-replicate t_i of the statistic t is created by randomly designating the members of the pair as first and second, eliminating the data from the first member of the pair, replacing the lost information with that from the second, re-poststratifying the weights, and then re-estimating the statistic based on this altered set of data.

The component of the sampling variability attributable to a pair of first-stage units is estimated as the squared difference between the value of the statistic for the complete sample and the pseudoreplicate associated with the pair. The estimated sample variance of the statistic t is the sum of the squared differences across the M pairs:

$$\hat{Var}(t) = \sum_{i=1}^{M} (t_i - t)^2 \quad .$$

In addition to having an estimator $\hat{Var}(t)$ of the sampling variance of a statistic t, it is important to have an indication of the number of degrees of freedom to attribute to estimated variance. The degrees of freedom of a variance estimator provide information on the stability of that estimator: the higher the number of degrees of freedom, the lower the variability of the estimator. In practical terms, the number of degrees of freedom of the variance estimator corresponds to the number of residual degrees of freedom that can be assumed for inferential procedures.

Empirical work at ETS on NAEP data (Johnson, 1989) has established that the degrees of freedom of a jackknife variance estimator is at most equal to the number M of paired first-stage sampling units and can be considerably less than that number. A procedure has been developed which estimates the effective number of degrees of freedom to assign to a jackknife variance estimator by considering the empirical distribution of design effect statistics. A design effect statistic is the ratio of the jackknife variance estimator to the conventional variance estimator based on simple random sampling assumptions. It can be shown that the approximate number of degrees of freedom to assign to jackknife variance estimates for some subpopulation corresponds to the degrees of freedom of the chi-square distribution which most closely matches the empirical distribution of design effect statistics for that subpopulation. The estimated degrees of freedom for selected key populations should be computed to provide a measure of the stability of the variance estimates.

1.2 Accounting for error of measurement

The previous discussion has considered the effects of the sample design on analysis when the variable of interest is observed without error on every respondent. This is the case for statistics such as weighted proportions-correct, but will not be the case for the subscale proficiency values. Because each individual will likely take only a few items within any subarea, that individual's subscale proficiency values will not be well determined. In order to account for this error of measurement in the NAEP, we have developed the plausible values methodology. It is important to realize that the variance of a statistic based on subscale proficiency values has two components: one due sampling variability which is estimable by the jackknife, the other due to the error of measurement in the subscale proficiencies of the sampled students. This latter component is estimated by assessing the dependence of the value of the statistic t on the particular set of student level estimated proficiencies used in its computation. The sum of these two components gives the proper estimate of the total amount of variability in the statistic t.

Specifically, let $T(\underline{\theta})$ be the value of some function of the vector $\underline{\theta}$ of unobservable proficiencies for each student. For example, $T(\underline{\theta})$ could be the mean proficiency across all students or it could be a more complex statistic such as a regression coefficient. It is important to recognize that there is error in the measurement of the proficiency values for any given student. That is, we do not know precisely the proficiency of any individual, but we can estimate the distribution of potential proficiency values for that individual. This estimated distribution takes into account the responses of the student to all items presented to that student as well as answers to certain background and demographic questions. A set of "plausible values" is created for each student by drawing values at random from the distribution of potential proficiency scale-scores for that student.

Let $\underline{\theta}_k^*$ be the vector consisting of the k^{th} plausible value for each of the sampled students. This vector is a plausible representation of what the vector $\underline{\theta}$ of true proficiencies might have been, had we been able to observe them. Let t_k be the value of $T(\cdot)$ evaluated using $\underline{\theta}_k^*$ as if those plausible

values were the true proficiencies. The statistic t_k provides a good estimate of the function $T(\underline{\theta})$, but does not individually indicate the effect on the estimate of not really knowing $\underline{\theta}$. An even better estimate of $T(\underline{\theta})$ is the average t^* of K (>1) such estimates t_k, where each of the estimates is based on a different set of the plausible values.

The component of the variance of t^* due to sampling variability is estimated by applying the jackknife variance estimator in turn to each of the statistics t_k, denoting the jackknife variance of t_k by U_k. The overall best estimate of the sampling variance of t^* is then the average, U^*, of the K sampling variances of the t_i.

To measure how well t^* approximates $T(\underline{\theta})$ given the error of measurement, we assess the dependence of the value of the statistic on the particular set of plausible values used in its computation. This "error due to imputation" is estimated by the variance among the K estimates:

$$B_K = \sum_{k=1}^{K} (t_k - t^*)^2 / (K-1).$$

The final estimate of the variance of t^* is the weighted sum of these two components:

$$V = U^* + (1 + K^{-1}) B_K \quad .$$

2. COMPARISON OF THE DATA FROM THE TRIAL STATE ASSESSMENT WITH THE DATA FROM THE NATIONAL ASSESSMENT

A major purpose of the Trial State Assessment is to allow the comparison of the levels of achievement of each of the participating states to the achievement level of the nation as a whole. A second major purpose is to enable comparisons among the participating states in terms of their achievement. The analyses of the data from the Trial State Assessment must be such that both of these purposes can be achieved.

An insightful analysis must take into account two characteristics of the Trial State Assessment which might affect the comparability of results between the states and between each state and the nation. First, since the within state assessments will be conducted in the field by state, district, and

school personnel, rather than by centralized field personnel as in the national NAEP, there is the possibility that the state level results might be affected by unforseen differences in administration procedures. This will affect the comparability of the results between the states to the extent that the effect of local administration varies by state. A consistent effect of local administration across the states will affect the comparability of each state's results with the results from the national NAEP.

There is a second difference between the Trial State Assessment and the national NAEP. We might expect there to be higher motivation on the part of students when their scores are to be used to compare their state with other states than in the case where the student's score is, at best, used for regional level reporting. If present, differential motivation between the state level and the national level assessments would certainly affect the comparability of the Trial State Assessment results to the national NAEP results. The comparisons among the states will also be affected to the extent that the motivation varies by state.

These differences between the Trial State Assessment and the national NAEP could have a notable effect on results. It has been our experience from the analysis of the National Assessment that relatively minor changes in administration procedures can have major impact on the measurement of subgroup proficiency measures, adding a component of variability to those measures which can be of greater magnitude than the subgroup differences that are typically considered significant. It is, accordingly, important to be able to detect, and correct for, any extraneous effects on achievement measures produced by the differences between the Trial State Assessment and the national NAEP. This detection is allowed for by two components of the 1990 assessment design:

(1) The national NAEP sample has been augmented to provide a subsample of schools which is representative of the aggregate of the states participating within the Trial State Assessment.

(2) A random half of the administrations within each of the participating states will be observed by central contractor staff for adherence to the required assessment procedures. In the course of this

observation, action will be taken to correct any deviations from these specified procedures.

Because of the corrective action taken in the event of problems, the local administration within the observed random half of the Trial State Assessment will be as close as is possible to the administration used by the national NAEP. This will reduce the differential effects due to administration procedures so that state level results based on the students participating in the observed administrations will be largely free of the confounding effects due to local rather than central contractor administration. It is reasonable to assume that the results from the observed and the unobserved local administrations will not differ in any consistent manner. Rather, it is likely that the major difference between these two types of administrations will be manifest as additional variance on the part of the unobserved administrations, due to the presence of potential problems in administration which would be resolved via corrective action taken during an observed administration.

Adopting this position leads to basing the analysis of the Trial State Assessment primarily on the (presumably less variable) data from the observed local administrations. In particular, the analyses of the relationship between the data from the Trial State Assessment and the data from the national NAEP should be based on the observed local administrations. How the data from the unobserved local administrations could be analyzed and how it might be combined with the data from the observed administrations will be discussed later.

It is important to note that the observation of the local administration will not remove the potential effect of differential motivation on the students. It is also possible that some of the effect of local administration will not have been eliminated through the process of observation and correction of observed problems. The average of these combined effects, averaging over the nation as a whole, can be detected by comparing the data from the observed local administrations with the appropriate subsample of the national NAEP.

2.1 Evaluation of the effect of type of administration

The State Aggregate Comparison (S.A.C.) subsample of the national NAEP consists of the grade eligible students selected from the schools in the augmented National Assessment sample. These students will be representative of the population of grade eligible students within the aggregation of the states which have agreed to participate in the Trial State Assessment so that the S.A.C. subsample and the Trial State Assessment will each provide nominally equivalent measures of achievement for random samples from the same population. Although the results from the two samples will differ somewhat due to sampling variability, any consistent difference may be attributable to effects of the type of assessment.

Any difference between the results for the Trial State Assessment and the results from the S.A.C. subsample of the national NAEP, will be manifested in one or more of the following ways:

(1) A consistent effect across all states for all items,

(2) An effect that depends on the type of item which is consistent across all states so that there is no effect of type of assessment for certain items, enhancing the scores from the Trial State Assessment for other items but decreasing the scores from the Trial State Assessment for yet other items.

(3) An effect which is consistent across all items but which depends on the type of state.

(4) An effect whose magnitude and direction depends both on the type of state and the type of item.

Each of the above effects of type of administration can be detected by a series of analyses performed on the item level achievement statistics. The results of those analyses will determine whether the Trial State Assessment results are directly comparable to the national NAEP results, and, if not, will indicate the type of equating that is required to produce comparable results.

Effects due to type of administration which are common across all states can be diagnosed by:

a) computing the percent correct statistics for each item separately for the aggregation of all states in the Trial State Assessment and for the S.A.C. subsample of the national NAEP,

b) converting these percent correct statistics into logits to eliminate floor and ceiling effects for very hard and very easy items,

c) plotting the logits from the aggregate state sample versus the corresponding logits from the S.A.C. subsample.

A consistent effect of type of administration which holds equally for all items will be indicated by a point cloud which clusters about a line with unit slope. The size of the effect in this case is indicated by the intercept of the line or, equivalently, under the assumption that the slope of the line is 1, by the mean difference between the logits for the two samples. If, upon comparison with the sampling variance, the value of this mean difference is not statistically significant, the data from the two samples can be considered comparable in their raw form. On the other hand, a significant difference would indicate that the scale score proficiency measures for the two samples should be equated. In this case, since the same relationship hold for all items, a common linear equating will be sufficient. A plot in which the points cluster along a line with a slope different from 1 has a more complicated interpretation but also indicates the need for, and sufficiency of, a common linear equating.

Other patterns in the plots of the logits can indicate the need for more refined equating procedures. For example, a plot in which the points follow a curvilinear path is contraindicative of linear equating and indicative of some more flexible alternative method such as equipercentile equating. As another example, a plot in which the points cluster along a number of distinct lines indicates a differential effect of type of assessment by class of item and further indicates the need for a different equating function for each of the distinct classes of items.

It is also possible to determine if the effect of the type of administration is differential by subgrouping of states. This analysis could be done by partitioning the states into classes (such as by region), computing the item level statistics within each of these classes for the aggregate state sample and for the S.A.C. subsample, and then comparing the results in the

same manner as above. (Note that it would not be possible to evaluate the effect of the type of administration for individual states since the S.A.C. subsample cannot provide reliable and representative results for most individual states. Further note that, although analyses to determine if the effect of type of administration appears differential by groupings of states are technically possible, they may not be practical due to the extreme time constraints placed on the production of the state level reports and to the political ramifications of proposing a different equating function for different collections of states.)

In addition to measuring the average effect of the type of administration across all the states, it is also useful to screen for individual items which function differentially between the two types of assessment. Since equivalent samples of students are being compared, differentially functioning items will appear as outliers on the plots of the logits, that is, as points which are unusually far away from the bulk of the point swarm.

A more refined screening for differentially functioning items can be obtained by conducting a series of Mantel-Haenszel (1959) analyses. Because of the focused-BIB structure (where each assessed student will be presented a booklet containing only a subset of the items), these analyses would have to be conducted separately by booklet. If the diagnostic plots indicated a differential effect of type of assessment by class of item (such as by content or process category) and there were a sufficient number of items available, the Mantel-Haenszel should be conducted separately for each of these defined classes.

Since the result of the Mantel-Haenszel analyses would be a large number of differential item functioning (DIF) statistics, one for each item, it would be extremely useful to have a method for screening this vast number of statistics. Such a method is the outlier detection procedure developed for the analysis of data from the International Assessment of Educational Progress (Johnson, 1989; Johnson and Tukey, 1987). This procedure compares the empirical distribution of the squares of the DIF statistics (divided by the squares of their standard errors) with the distribution to be expected under the null hypothesis of no differential item functioning. The procedure controls for the effects of multiple tests, discounting DIF statistics which

are no larger than bounds based on what would be expected in the case of a large number of tests, and flagging as outliers DIF statistics which exceed those bounds. Items designated as outliers by this procedure should be treated as different within the two samples.

2.2 Equating the observed administration results to the national results

It is likely that a linear equating will be adequate to link the results of the Trial State Assessment to the national NAEP if, indeed, any is needed. If warranted by the diagnostic plots, the Trial State Assessment results could be put onto the same scale as the national NAEP via a linear equating achieved by the following steps:

a) Estimate the item parameters, by subscale, for the full national NAEP sample. This item calibration should include all students from all age classes involved in the national assessment of the subject area. This is important for the estimation of parameters for items which are presented to two age classes and which are at the extremes of the difficulty continuum for at least one of the age classes. Parameter estimates for such items based on only one of the two age classes tend to be unstable.

b) Using the item parameters from (a), obtain subscale and composite plausible values for all students assessed as a part of the Trial State Assessment.

c) Compute the mean and standard deviation of the plausible values of all students assessed in an observed administration in the aggregation of the states in the trial assessment. In this computation, the students must be weighted to represent the target population based on the aggregation of the states. The trial state results can then be linked to the national NAEP results by matching this mean and standard deviation to the mean and standard deviation of all students in the S.A.C. subsample of the national NAEP.

If the diagnostic plots indicated the need for linear equating by subscale, the matching of moments in step (c) above should be by subscale. If the plots indicated that linear equating is not adequate, then alternative methods of equating, such as equipercentile equating, must be employed.

2.3 Linking the unobserved administration results to the national results

Whatever equating function that is used to link the data from the observed local administration component of the Trial State Assessment to the national NAEP should also be applied to the results from the unobserved administrations. Although it is in principle possible to evaluate the effect of observed versus unobserved administration on a state by state basis, it should be pointed out that the tight time constraints placed on the reporting of state level results, as well as the political ramifications of potentially discounting the unobserved administration data for some states or proposing different equating functions for each state, council against such an analysis.

It is possible to avoid appearing judgmental of the data from the unobserved administrations while, at the same time, providing state level data of the highest possible quality by providing three sets of data and statistics for each state:

1) data and statistics based only on the observed administrations, the state level data which is as close as is possible in administration to the national data,
2) data and statistics based only on the unobserved administrations but equated to the national data using the equating functions for the observed administration, and
3) data and statistics based on the combination, with equal weights, of the results from the observed and unobserved administrations.

3. BETWEEN STATE COMPARISONS

Perhaps the major interest from the Trial State Assessment will be in the opportunity to compare the various states to each other in terms of measures of the educational performance of their students. As noted in the report of the NAEP planning project under the auspices of the Council of Chief State School Officers (CCSSO, 1988), the comparison of the states to each other will be inherently judgmental. Because of this, it is not appropriate to merely present state level results without accompanying information to establish a

context for comparison. As the CCSSO points out, it is important to acknowledge that states differ in the economic and demographic settings in which the schools operate and in the social and economic background of the students entering those schools. Certain of these variables might indicate difficulties faced by schools as they educate their particular populations and, to the extent that the distribution of these variables differs by state, may account for some part of different outcomes among states. It should be noted that the same issues arise in the comparison of each of the states to the nation as a whole.

There are a number of approaches which might be used to account for differences in the characteristics in the populations within the states, among these being:

- the reporting of proficiency levels for standardized populations,
- the comparison of each state's raw score to a predicted score based on the mix of students in the state,
- the reporting of raw results using a rank-ordered scheme in which the states are ranked according some relevant measure, and
- the reporting of raw results by state separately by similar type of student.

The first of these methods divides the students within each state into subgroups which are then recombined according to the relative proportion of each of the subgroups in a standardized population generally corresponding to the full United States population. Since the state level subgroup proficiency values are held fixed at their observed values, this reweighting corresponds to a prediction of how each state would do if it had the same subgroup distribution as in the reference population but the members of each subgroup performed as they were observed to do in the state.

The second of these methods, which is the converse of the first, holds the state results fixed and presents a comparison based on a combination of subgroup scores for the reference population using the relative proportions of each of the subgroups observed in the state.

Neither the first nor the second method directly addresses the comparisons of the states to each other. The first method allows the comparison of

adjusted scores between the states, but the utility of the comparison depends on the subgroups used for adjustment. Furthermore, the use of adjusted scores might be viewed with some suspicion by the general public. The second method retains the raw scores for each state and provides some indication of what the expected scores might be if there were a different demographic mix, but does not directly allow the comparison of the states to each other.

The third and fourth methods address the comparability of the states, in terms of their observed scores. As an example of the third method, the NAEP planning project of the CCSSO proposed that the states be rank ordered on a composite of state demographic variables. They suggested that three indicators which appear to provide an appropriate indication of the differences between the states in terms of their general population are: per capita income, the percent of adults who have completed four years of high school and the percent of children who live in poverty. Other variables could be considered in addition to or instead of these variables. The key constraint is that the variable represent characteristics of the general population which is not under the direct control of the state or schools and which plausibly might be related to the performance of the students. The procedure is to then determine a composite index which will allow the ranking of the states along some continuum which, in the current example, might be labelled socioeconomic status. In this ranking, states with similar values on the index will presumably have similar amounts of resources and face similar difficulties in the education of their students.

The achievement of each state could then be compared to the achievement of the states which appear immediately above and below that state in the ordered list, the index allowing comparability by approximately controlling for the effects of whatever variables were used in its construction. It is possible to argue that the variables should be limited to nonmodifiable or out-of-school variables on the grounds that these represent the inputs that the schools have to work with. However, as pointed out by Haertel(1988), these out-of-school variables can also adjust for the effects of school practice and policy variables to the extent that the out-of-school variables are correlated with the in-school variables.

In addition to developing comparisons among states which take into account factors not under the control of the schools within the state, it would also be of great interest to consider the impact of factors which are directly under the control of the schools and which would plausibly have an effect on the education of the students. Examples of these are factors related to the background and training of the teachers and their instructional practices; the per pupil expenditures; and the particular curriculum actually used within the schools. For example, in the 1990 Trial State Assessment, mathematics coordinators and teachers will be asked to provide indications of the percentages of students within their school which might be expected to have had the opportunity to learn the material included in the assessment. A ranking of the states on the basis of an opportunity to learn index would tend to control for differences between the states in curricula, so that adjacent states (in terms of the ranking) would tend to have similar curricular concentrations.

Given a selected set of variables, the ranking index could be constructed either via a regression in which achievement is related to the background variables and the best predicting index selected or via a principal component analysis in which the index which maximally discriminates between the states without regard to its effect on performance is selected.

An alternative way to compare the states, also suggested by the CCSSO NAEP planning project, is based on the background of the students being assessed. In this technique, the background characteristics of the students within each state are considered. The aim is to construct groupings of students within each state in terms of background characteristics not under the control of the schools but known to be related to performance. Examples of such variables are the levels of education of the parents and the number of home study items such as books, encyclopedias and periodicals. After constructing such groupings, presumably similar students in different states can be compared. To the extent that the groupings defined are, in fact, comparable across states, this will provide fair between-state comparisons. The chief difficulties with this approach are the possible inadequate sample sizes for some groupings within some states and the potential need to create a large number of groupings to ensure comparability.

It is, of course, possible and probably desirable to employ a number of different procedures for between-state comparisons. Each technique will provide its own insights into how differences between the states are related to differences in achievement.

SUMMARY

The Trial State Assessment will provide a rich source of information about the knowledge, skills, understandings, backgrounds and attitudes of students within the states which choose to participate. We have considered a number of the statistical issues which must be addressed in the analyses of these data in order to yield useful state level results. It was observed that scaling models should be used to produce estimates of the distribution of abilities within the subject area, that these scaling models should distinguish between content areas, and that the models require special technology to properly take the amount of available information into account. It was further observed that the structure of the samples must be taken into account in analysis. The importance of attending to the potential effects of differences in administration procedures and differences in motivation was stressed. And, finally, it was noted that the fact that the states differ in terms of demographic, socioeconomic and other indicators is important, and that these differences may, at least in part, affect the comparability of achievement results between states. Analyses which consider all of these issues will produce state level results which fairly and accurately reflect the condition of education within the states and which provide some insight into where instructional improvement efforts should be focused.

References

Council of Chief State School Officers. (1988, March). On reporting student achievement at the state level by the National Assessment of Educational Progress (Report of the National Assessment Planning Project, Wilmer S. Cody, director). Washington, DC: Author.

Haertel, E. (1988). Within-state comparisons: suitability of state models for national comparisons. In Edward Haertel, Chair, Report of the technical review panel on the National Assessment of Educational Progress. Washington, DC: National Center for Educational Statistics.

Johnson, E. G. (1988). Considerations and techniques for the analysis of NAEP data. (ETS Research Report, RR-88-49) Princeton, NJ: Educational Testing Service.

Johnson, E. G. (1989). The application of outlier detection to DIF analysis. Princeton, NJ: Educational Testing Service.

Johnson, E. G. & Tukey, J. W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In Collin Mallows (Ed.), Design, data, and analysis. New York: Wiley & Sons.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.